# From MARC silos to Linked Data silos?

Suominen, Osma Ilmari

2017

Suominen , O I & Hyvönen , N K 2017 , ' From MARC silos to Linked Data silos? ' , o-bib.
Das offene Bibliotheksjournal , vol. Bd. 4, (2017) , no. Nr. 2 . https://doi.org/10.5282/o-bib/2017H2S1-13

http://hdl.handle.net/10138/235981
https://doi.org/10.5282/o-bib/2017H2S1-13

## Aufsätze

# From MARC silos to Linked Data silos?

*Osma Suominen, National Library of Finland*
*Nina Hyvönen, National Library of Finland*

**Summary:**
Libraries are opening up their bibliographic metadata as Linked Data. However, they have all used different data models for structuring their bibliographic data. Some are using a FRBR-based model with several layers of entities while others use flat, record-oriented data models. The proliferation of data models limits the reusability of bibliographic data. In effect, libraries have moved from MARC silos to Linked Data silos of incompatible data models. Data sets can be difficult to combine and reuse. Small modelling differences may be overcome by schema mappings, but it is not clear that interoperability has improved overall. We present a survey of published bibliographic Linked Data, the data models proposed for representing bibliographic data as RDF, and tools used for conversion from MARC. Also, the approach of the National Library of Finland is discussed.

**Zusammenfassung:**
Seit einiger Zeit stellen Bibliotheken ihre bibliografischen Metadadaten verstärkt offen in Form von Linked Data zur Verfügung. Dabei kommen jedoch ganz unterschiedliche Modelle für die Strukturierung der bibliografischen Daten zur Anwendung. Manche Bibliotheken verwenden ein auf FRBR basierendes Modell mit mehreren Schichten von Entitäten, während andere flache, am Datensatz orientierte Modelle nutzen. Der Wildwuchs bei den Datenmodellen erschwert die Nachnutzung der bibliografischen Daten. Im Ergebnis haben die Bibliotheken die früheren MARC-Silos nur mit zueinander inkompatiblen Linked-Data-Silos vertauscht. Deshalb ist es häufig schwierig, Datensets miteinander zu kombinieren und nachzunutzen. Kleinere Unterschiede in der Datenmodellierung lassen sich zwar durch Schema Mappings in den Griff bekommen, doch erscheint es fraglich, ob die Interoperabilität insgesamt zugenommen hat. Der Beitrag stellt die Ergebnisse einer Studie zu verschiedenen veröffentlichten Sets von bibliografischen Daten vor. Dabei werden auch die unterschiedlichen Modelle betrachtet, um bibliografische Daten als RDF darzustellen, sowie Werkzeuge zur Erzeugung von entsprechenden Daten aus dem MARC-Format. Abschließend wird der von der Finnischen Nationalbibliothek verfolgte Ansatz behandelt.

**Citable Link (DOI):** http://dx.doi.org/10.5282/o-bib/2017H2S1-13
**Author Identification:** Osma Suominen: ORCID http://orcid.org/0000-0003-0042-0745;
Nina Hyvönen: ORCID http://orcid.org/0000-0001-6467-5961
**Keywords:** Bibliographic metadata, Linked data, Open data, Data models, FRBR, RDF

## 1. Introduction

Many libraries are experimenting with publishing their bibliographic metadata as Linked Data. The stated purpose of such experiments is to open up bibliographic silos, typically based on MARC (Machine Readable Cataloging) records, into the wider world and make them more interoperable, accessible and understandable for developers who are not intimately familiar with library data.

The libraries who have so far published bibliographic metadata as Linked Data have all followed different data models, schemas and/or ontologies for structuring their data. Most of these published data sets originate in MARC records. Some libraries have opted to transform their records into a FRBR (Functional Requirements for Bibliographic Records) oriented model where Works, Expressions and Manifestations are represented as separate entities. Others have simply performed a conversion from MARC to Dublin Core, dumbing down their records into a lowest common denominator format. There is currently no universal model for how to represent bibliographic metadata as Linked Data, even though many attempts for such a model have been made.

We present a survey of the landscape of published bibliographic Linked Data and the various tools used for conversion from MARC records. We also analyze the data models, schemas and ontologies that have been proposed for representing bibliographic data as Linked Data.

While Linked Data promises smoother interoperability between different data sets, the proliferation of data models for bibliographic data represents a significant barrier to reusing the data. In effect, the library community has moved from MARC silos to Linked Data silos of incompatible data models. Bibliographic data expressed using different models is difficult to combine and use together. While small differences in modelling may be overcome by ontology mappings and/or translation mechanisms, it is not clear that interoperability has improved overall. It is obvious, however, that libraries need a shared vision and strategy on how to open up their data sustainably.

At the National Library of Finland, we have already taken steps to open up authority data, including the multilingual General Finnish Ontology YSO[1], as Linked Data via the Finto service[2]. We are following up that work by also opening up bibliographic metadata as Linked Data while trying to learn from the examples of others. We aim at building a Linked Data solution that makes bibliographic data more accessible and interoperable instead of adding yet another incompatible data model to the current mix.

1 Satu Niininen, Susanna Nykyri and Osma Suominen, „The Future of Metadata: Open, Linked, and Multilingual – the YSO Case," *Journal of Documentation* 73.3 (2017): 451-465, http://dx.doi.org/10.1108/JD-06-2016-0084.

2 Osma Suominen et al., „Deploying National Ontology Services: From ONKI to Finto," (paper presented at the International Semantic Web Conference (Industry Track), Riva del Garda, Italy, October 19-23, 2014), accessed May 24, 2017, http://www.ceur-ws.org/Vol-1383/paper6.pdf.

## 2. Current bibliographic data models

In this section we present a survey of the main bibliographic data models that have been used for publishing bibliographic Linked Data, as well as some conversion tools and bibliographic Linked Data sets. Many of these data models and data sets have also been described in a 2011 W3C Library Linked Data incubator report.[3] Hillman, Dunsire and Phipps discuss the evolutionary pressures faced by bibliographic data models, for example the inherent tension between simplicity and semantic accuracy.[4]

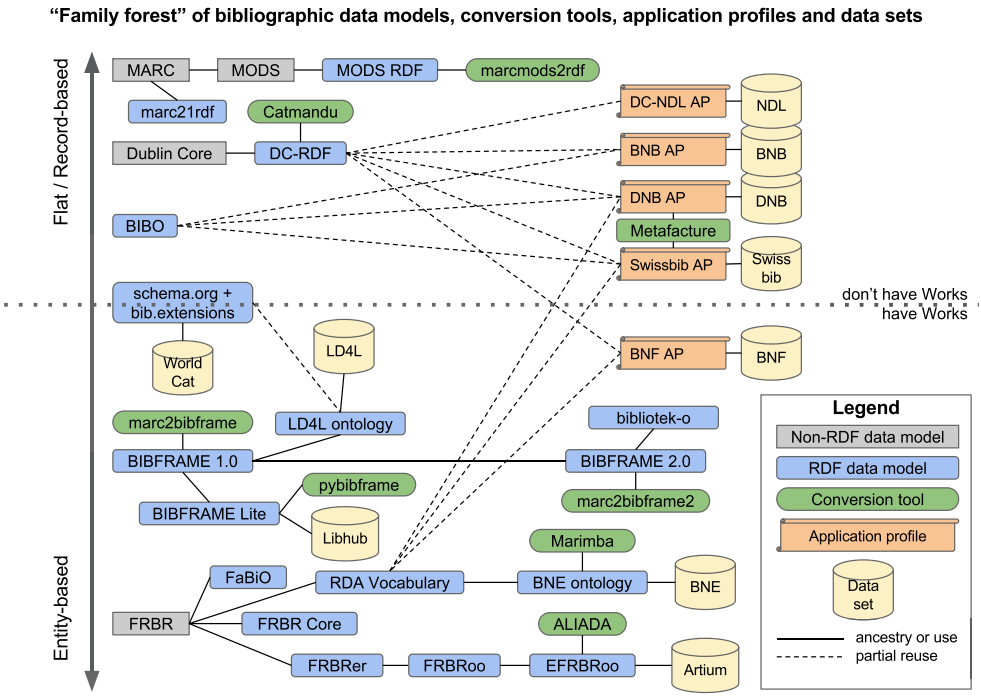


*Figure 1: Family forest of bibliographic data models*

An overview of the bibliographic data landscape is shown in Figure 1, which also traces the ancestry of some of the data models in the form of family trees. Collectively they form a „family forest", since there is no common ancestor for all the models. On the vertical axis, the data models are ordered based on the specificity of modelling: flat, record-oriented models that don't distinguish between

3 Antoine Isaac et al., „Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets," in *W3C Incubator Group report* (2011), accessed May 24, 2017, http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/.

4 Diane I. Hillmann, Jon Phipps and Gordon Dunsire, „Maps and Gaps: Strategies for Vocabulary Design and Development," (paper presented at the International Conference on Dublin Core and Metadata Applications 2013, Lisbon, Portugal, September 2-6, 2013), accessed May 24, 2017, http://dcpapers.dublincore.org/pubs/article/view/3673/1896.

 

layers of bibliographic entities are at the top, while the most detailed models are near the bottom. A dotted horizontal line separates the models that don't represent Works as separate entities from those that do. Non-RDF data models, such as XML schemas, are represented as gray boxes and RDF-based data models as blue boxes. Conversion tools are shown as green boxes, with lines connecting them to the output data model that they support. Application profiles are represented as orange scrolls and published Linked Data sets as yellow cylinders. Solid lines connect closely related data models, while dashed lines represent partial reuse of data model elements.

## 2.1. Flat, record-based models

MARC is the original record-based data model for bibliographic data. There have been a few attempts to produce an RDF version of MARC, but the impedance mismatch (i.e., differences in modelling style and structural principles) between MARC structures and RDF generally makes that approach difficult. One attempt at an RDF version of MARC is marc21rdf.[5]

MODS (Metadata Object Description Schema)[6] is an XML schema for representing bibliographic data, which can represent a large subset of the data that can be encoded in a MARC bibliographic record. An RDF version of MODS is called MODS RDF[7] and the marcmods2rdf[8] tool can convert from MARC via MODS to MODS RDF.

Dublin Core (including the original DC Elements[9] and the newer DC Terms[10]) is an abstract metadata model oriented around bibliographic data published on the Web. DC-RDF[11] is the recommended way of expressing DC metadata using RDF. The Catmandu[12] tool can, among many other possible uses, convert metadata from MARC to DC-RDF.[13,14]

BIBO (Bibliographic Ontology)[15] is an OWL ontology for expressing bibliographic information. It is mainly oriented around academic publishing and represents the kind of data needed for creating citations and lists of references for scientific papers.

5 „The MARC21 Vocabularies," Metadata Management Associates, accessed May 24, 2017, http://www.marc21rdf.info/.
6 „Metadata Object Description Schema Official Web Site," Library of Congress, accessed May 24, 2017, http://www.loc.gov/standards/mods/.
7 „MODS RDF Ontology", Library of Congress, accessed May 24, 2017, https://www.loc.gov/standards/mods/modsrdf/.
8 „Converting MARC or MODS to RDF with the Simile Stylesheets", accessed May 24, 2017, https://github.com/cul/marcmods2rdf.
9 „Dublin Core Metadata Element Set, Version 1.1", Dublin Core Metadata Initiative, 14 June 2012, accessed May 24, 2017, http://dublincore.org/documents/dces/.
10 „DCMI Metadata Terms", Dublin Core Metadata Initiative, 14 June 2012, accessed May 24, 2017, http://dublincore.org/documents/dcmi-terms/.
11 „Expressing Dublin Core Metadata Using the Resource Description Framework (RDF)", Dublin Core Metadata Initiative, 14 January 2008, accessed May 24, 2017, http://dublincore.org/documents/dc-rdf/.
12 „Catmandu: a Data Toolkit", LibreCat, accessed May 24, 2017, http://librecat.org/Catmandu/.
13 Christina Harlow, „Data Munging Tools in Preparation for RDF: Catmandu and LODRefine," *Code4Lib Journal* 30 (2015), accessed May 24, 2017, http://journal.code4lib.org/articles/11013.
14 „Day 15: MARC to Dublin Core," LibreCat (blog), 19 December, 2014, https://librecatproject.wordpress.com/2014/12/19/day-xx-marc-to-dublin-core-12/.
15 Frédérick Giasson and Bruce D'Arcus, „Bibliographic Ontology Specification," 4 November, 2009, accessed May 24, 2017, http://bibliontology.com/.

## 2.2. Hybrid models

Schema.org[16] is a general vocabulary and set of practices for describing many kinds of structured data on web pages, including bibliographic information. The bibliographic extensions[17] of Schema.org, some of which have been incorporated to the core Schema.org model, provide a mechanism for separating different levels of entities such as works and instances/manifestations, but simple Schema.org structures can also be used for representing flat bibliographic data similar to DC or BIBO.

OCLC has been publishing WorldCat Linked Data[18] modelled according to Schema.org since 2012. Originally the data model was flat, as the data was created by direct conversion of WorldCat MARC records. In 2014, OCLC released the WorldCat Works[19] data set, which provides an additional layer of work entities based on an analysis of the MARC records. It is modelled using the Schema.org bibliographic extensions.

## 2.3. BIBFRAME family of entity-based models

BIBFRAME (Bibliographic Framework)[20] is an initiative by the Library of Congress for providing a new, RDF-based foundation for bibliographic description. From a modelling perspective, one of the main features of BIBFRAME is the separation between Works and Instances. The BIBFRAME 1.0 version, released in 2014, was the starting point for further development of the standard. The Library of Congress also developed the marc2bibframe[21] conversion tool that can convert MARC records into BIBFRAME 1.0 compliant RDF.

Zepheira, the company that worked with the Library of Congress on BIBFRAME 1.0, subsequently produced their own, modular set of RDF vocabularies, with the BIBFRAME Lite[22] model at its core. While these vocabularies share much of the BIBFRAME 1.0 abstract modelling, they do not use the same RDF classes or properties. Zepheira has also provided a conversion tool called pybibframe[23] for converting MARC records into their model. The Libhub Initiative[24], spearheaded by Zepheira, aims at publishing bibliographic resources to the Web using the BIBFRAME Lite model.

The Linked Data for Libraries (LD4L) project[25] converted Cornell, Harvard and Stanford university library data sets to BIBFRAME 1.0. In the process, these libraries found some BIBFRAME structures awkward and replaced them with structures from other RDF vocabularies, including Schema.org,

16 „schema.org," accessed May 24, 2017, http://schema.org/.
17 „Schema.org Hosted Extension: bib," accessed May 24, 2017, https://bib.schema.org/.
18 „OCLC Linked Data," OCLC, accessed May 24, 2017, https://www.oclc.org/developer/develop/linked-data.en.html.
19 „WorldCat Work Descriptions," OCLC, accessed May 24, 2017, https://www.oclc.org/developer/develop/linked-data/worldcat-entities/worldcat-work-entity.en.html.
20 „Bibliographic Framework Initiative," Library of Congress, accessed May 24, 2017, https://www.loc.gov/bibframe/.
21 „XQuery Utility to Convert MARC/XML Bibliographic Records to BIBFRAME Resources," accessed May 24, 2017, https://github.com/lcnetdev/marc2bibframe.
22 „BIBFRAME Lite + Supporting Vocabularies," Zepheira, accessed May 24, 2017, http://bibfra.me/.
23 „Some Open-source Tools for Working with BIBFRAME," accessed May 24, 2017, https://github.com/zepheira/pybibframe.
24 „Join the Movement - Take the Libhub Initiative Pledge," Zepheira, accessed May 24, 2017, http://www.libhub.org/.
25 „Linked Data for Libraries: the Gateway," LD4L, accessed May 24, 2017, https://www.ld4l.org/.

 

and created the LD4L Ontology.[26] The project has also published draft LD4L data sets[27] using this data model, created from the catalogs of the university libraries.

BIBFRAME 2.0[28] was released by the Library of Congress in April 2016. Some aspects of the LD4L critique of BIBFRAME 1.0 were considered in developing the new standard. The marc2bibframe2[29] conversion tool that can convert bibliographic data from MARC to BIBFRAME 2.0 was released in March 2017 by the Library of Congress and the company Index Data.

The Linked Data for Production (LD4P) and Linked Data for Libraries-Labs (LD4L-Labs) projects, which succeeded LD4L, convert bibliographic data sets to BIBFRAME 2.0, but extend the data model as necessary, aligning it with other established RDF data models. The resulting model bibliotek-o[30] will thus be based on BIBFRAME 2.0, but with some differences from the original. The details of the model are still being defined. The LD4L-Labs project has also started work on a conversion tool[31] from MARC to their target data model as well as a validation tool[32] to assess the output of different MARC to BIBFRAME conversion tools.

## 2.4. FRBR family of entity-based models

FRBR is an abstract, conceptual model for bibliographic data developed by IFLA. The FRBR model is known for its separation between Works, Expressions, Manifestations and Items, collectively known as the WEMI model. FRBR Core[33] was an early attempt at defining an RDF vocabulary for representing FRBR entities, while FRBRer[34] can be considered the official RDF representation of FRBR as it was produced by the IFLA FRBR study group. The newest incarnation of FRBR is the Library Reference Model[35] (FRBR-LRM), which aims to be a consolidated high-level conceptual reference model that covers all aspects of bibliographic data.

FaBiO (FRBR-aligned Bibliographic Ontology)[36] is a FRBR-based OWL ontology for representing scholarly publications. It is similar in scope to BIBO, but the modelling is based on the WEMI model and thus allows describing different levels of bibliographic entities.

26 „LD4L 2014 Ontology," LD4L, accessed May 24, 2017, https://www.ld4l.org/ontology.
27 „Downloads," LD4L, accessed May 24, 2017, http://draft.ld4l.org/downloads/.
28 „BIBFRAME Model, Vocabulary, Guidelines, Examples, Notes, Analyses," Library of Congress, accessed May 24, 2017, https://www.loc.gov/bibframe/docs/index.html.
29 „marc2bibframe2: Convert MARC Records to BIBFRAME2 RDF," Library of Congress, accessed May 24, 2017, https://github.com/lcnetdev/marc2bibframe2.
30 „bibliotek-o: a BIBFRAME 2 Extension Ontology", accessed May 24, 2017, http://bibliotek-o.org/.
31 „bib2lod: Converts Bibliographic Records to Linked Open Data," accessed May 24, 2017, https://github.com/ld4l-labs/bib2lod.
32 „Marc to Bibframe Validation," accessed May 24, 2017, https://github.com/ld4l-labs/marc2rdf-validator.
33 Ian Davis, Richard Newman and Bruce D'Arcus, „Expression of Core FRBR Concepts in RDF," 2005, accessed May 24, 2017, http://vocab.org/frbr/core.
34 „FRBRer model," accessed May 24, 2017, http://metadataregistry.org/schema/show/id/5.html.
35 „World-wide Review of the FRBR-Library Reference Model, a Consolidation of the FRBR, FRAD and FRSAD Conceptual Models", IFLA, 28 February, 2016, accessed May 24, 2017, https://www.ifla.org/node/10280.
36 David Shotton, Silvio Peroni, Paolo Ciccarese and Tim Clark, „FaBiO, the FRBR-aligned Bibliographic Ontology," 11 July 2016, accessed May 24, 2017, http://www.sparontologies.net/ontologies/fabio/source.html.

FRBRoo (FRBR Object Oriented)[37] is an object-oriented interpretation of FRBR and FRBRer. It is developed as an extension of the CIDOC CRM conceptual model for cultural heritage and has been published as an RDF Schema. EFRBRoo (Erlangen FRBRoo)[38] is an OWL ontology for expressing FRBRoo data. The ALIADA project[39] has developed a conversion and publishing tool that can convert MARC bibliographic records to RDF data based on the EFRBRoo model. The ALIADA project partners have published metadata using the tool; for example, the Basque museum center for contemporary art Artium has published a dataset as Linked Data.[40]

RDA (Resource Description and Access) is a standard for descriptive cataloguing of bibliographic data which is organized based on the FRBR WEMI model. The RDA Vocabularies are a set of RDF classes and properties, published via the RDA Registry[41] web site, which can be used to represent bibliographic data catalogued according to the RDA rules.

The National Library of Spain (BNE), together with the Ontology Engineering Group at Universidad Politécnica de Madrid, has defined the BNE Ontology[42] for BNE bibliographic RDF data. The ontology defines classes and properties, many of which are derived from the RDA Vocabulary as well as other RDF vocabularies. BNE defined their own data model instead of using existing models because they wanted greater control over the data model.[43] The Marimba[44] tool was used to convert the BNE MARC records into RDF. The resulting BNE RDF data has been published as Linked Data.[45]

## 2.5. Application profiles

An application profile is a document that specifies how metadata elements from existing data models, possibly including locally defined additions, are combined and reused for a particular application. It can be expressed as a technical document, a machine-readable schema or just a consistently applied informal set of conventions.

The National Diet Library of Japan (NDL) has defined the DC-NDL application profile,[46] which specifies how DC elements as well as NDL specific additions are used in Japanese bibliographic metadata

37 „Definition of FRBRoo: A Conceptual Model for Bibliographic Information in Object-Oriented Formalism," IFLA, last updated May 22, 2017, accessed May 24, 2017, https://www.ifla.org/publications/node/11240.
38 „Erlangen Functional Requirements for Bibliographic Records Object-oriented (EFRBRoo)," University of Erlangen-Nuremberg, accessed May 24, 2017, http://erlangen-crm.org/efrbroo.
39 „ALIADA: Automatic Publication under Linked Data Paradigm of Library Data," ALIADA Consortium, accessed May 24, 2017, http://www.aliada-project.eu/.
40 „Artium Aliada Dataset," Artium, accessed May 24, 2017, http://aliada.scanbit.net:8891/.
41 „RDA Registry," accessed May 24, 2017, http://www.rdaregistry.info/.
42 „BNE Ontology," Biblioteca Nacional de España, accessed May 24, 2017, http://datos.bne.es/def/ontology.html.
43 „Data Model," Biblioteca Nacional de España, accessed May 24, 2017, http://www.bne.es/en/Inicio/Perfiles/Bibliotecarios/DatosEnlazados/Modelos/.
44 „Marimba: Unlocking Your Library Data," accessed May 24, 2017, http://marimba4lib.com/.
45 „El portal de datos bibliográficos," Biblioteca Nacional de España, accessed May 24, 2017, http://datos.bne.es/.
46 „RDF Schema Declaration for NDL Metadata Terms," National Diet Library, last modified 1 December, 2011, http://ndl.go.jp/dcndl/terms/.

 

published as Linked Data.[47] Their main concerns in defining this application profile were representing Japanese linguistic features and providing specialized metadata for digitized materials.[48]

The British Library has produced application profiles for representing books and serials in the British National Bibliography (BNB) as Linked Data.[49] Both are based on reuse of DC, BIBO, FOAF (Friend of a Friend),[50] the ISBD element set[51] and other RDF vocabularies. The British Library decided not to use a FRBR based data model, since it would have required a large investment to identify WEMI entities in the existing MARC records.[52]

The German National Library (DNB) has defined an application profile[53] for bibliographic RDF data that combines DC, FOAF, ISBD and RDA elements, Schema.org bibliographic elements and other RDF vocabularies. This model is used in the data published via the DNB Linked Data Service.[54] DNB relies on the Metafacture[55] toolkit for conversion of their MARC-based data into RDF.

Swissbib, a union catalog for Swiss libraries, has defined a data model[56] for the Swissbib Linked Data service[57] that combines DC, BIBO, FOAF, the RDA vocabularies as well as other RDF vocabularies. Like DNB, Swissbib uses the Metafacture toolkit for conversion of MARC records into RDF.

The French National Library (BnF) has defined their own data model[58], reusing existing RDF vocabularies, including DC and some parts of the RDA Vocabulary, to create a BnF Application Profile, which is structured according to the FRBR/RDA WEMI model. The BnF data set is available as Linked Data.[59]

47 „Use and Connect: What is NDL Linked Open Data (LOD)?," National Diet Library, accessed May 24, 2017, http://ndl.go.jp/en/aboutus/standards/lod.html.

48 Saho Yasumatsu, Akiko Hashizume and Julie Fukuyama, „National Diet Library Dublin Core Metadata Description (DC-NDL): Describing Japanese Metadata and Connecting Pieces of Data," (paper presented at the International Conference on Dublin Core and Metadata Applications 2016, Copenhagen, Denmark, October 13-16, 2016), accessed May 24, 2017, http://dcevents.dublincore.org/IntConf/dc-2016/paper/view/437/485.

49 „Free Data Services," British Library, accessed May 24, 2017, http://www.bl.uk/bibliographic/datafree.html.

50 „FOAF Vocabulary Specification 0.99," 14 January, 2014, accessed May 24, 2017, http://xmlns.com/foaf/spec/.

51 „ISBD elements," Open Metadata Registry, accessed May 24, 2017, http://metadataregistry.org/schema/show/id/25.html.

52 Corine Deliot, „Publishing the British National Bibliography as Linked Open Data," *Catalogue & Index* 174 (2014): 13-18. http://www.bl.uk/bibliographic/pdfs/publishing_bnb_as_lod.pdf.

53 „The Linked Data Service of the German National Library: Modelling of Bibliographic Data," Deutsche Nationalbibliothek, 13 September, 2016, accessed May 24, 2017, http://www.dnb.de/SharedDocs/Downloads/EN/DNB/service/linkedDataModellierungTiteldaten.pdf?__blob=publicationFile.

54 „Linked Data Service of the German National Library," Deutsche Nationalbibliothek, accessed May 24, 2017, http://www.dnb.de/EN/Service/DigitaleDienste/LinkedData/linkeddata_node.html.

55 „Metafacture-core Wiki," Culturegraph Project, accessed May 24, 2017, https://github.com/culturegraph/metafacture-core/wiki.

56 „Swissbib Data Goes Linked Teil 1: Metadatentransformation, Modellierung, Indexierung," *swissbib-info* (blog), 13 April, 2016, accessed May 24, 2017, http://swissbib.blogspot.fi/2016/04/swissbib-data-goes-linked-teil-1.html.

57 „Swissbib Search Home," Swissbib, accessed May 24, 2017, http://linked.swissbib.ch/.

58 „Semantic Web and data model," Bibliothèque nationale de France, accessed May 24, 2017, http://data.bnf.fr/en/semanticweb.

59 „data.bnf.fr," Bibliothèque nationale de France, accessed May 24, 2017, http://data.bnf.fr/.

## 3. Use cases for bibliographic data

The choice of data model as well as modelling choices within a data model are affected by the intended use case of the data. We can make a rough distinction between two ends on a spectrum of use cases: Maintaining library data natively as RDF, or publishing Linked Data on the Web for others to reuse. We call these two extremes *libraryish* and *webbish* use cases, respectively. As Hillmann, Dunsire and Phipps note, „the ontology you use to maintain your metadata never has to be the ontology you use to publish your metadata".[60]

### 3.1. Libraryish use cases

In the *libraryish* use case, bibliographic data is produced and maintained natively as RDF triples. When converting data from legacy formats such as MARC, it is very important to preserve as much as possible of the details in the original data, as it will otherwise be unavailable for further maintenance.

Converting all the detailed information implies that so-called „housekeeping metadata" such as timestamps and status information has to be represented in the RDF data. It often also means that the modelling is on a more abstract level: instead of modelling real world objects, the RDF model may end up modelling another level of abstraction that corresponds to the legacy records. For example, people are not modelled as human beings, but as records in a name authority list.

When data is maintained natively as RDF, there are many specific requirements for the data model. In practice, these requirements make it more likely that the data model will be self-contained, with little reuse of other data models. A self-contained data model gives greater control to the organization defining it, but the lack of reuse makes interoperability with other data models more difficult.

### 3.2. Webbish use cases

In a *webbish* use case, bibliographic data is shared on the Web as Linked Data for others to reuse. The original data is generally maintained not as RDF, but in legacy formats such as MARC.

Important goals in this kind of setting include the interoperability of data with other data sets and following best practices for Linked Data publishing such as providing resolvable URIs, a SPARQL endpoint and bulk downloads. In data modelling, simplicity is favored over exhaustive detail and some loss of information is acceptable, if it makes the publishing and consuming of data easier. Entities such as places, people and organizations are typically modelled as real world objects, not as authority records.

### 3.3. Existing data models by use case

In Figure 2, we have placed some RDF data models on a spectrum between *libraryish* and *webbish* use cases. In the top area are models which are suitable for bibliographic data (i.e. roughly what is represented in a MARC bibliographic record), the bottom area contains data models which are suitable for representing auxiliary entities such as people, concepts and places (which could be represented

60 Hillmann, Dunsire and Phipps, „Maps and Gaps", 88.

 

as MARC authority records), and the middle area contains models that can represent both kinds of information. Arrows connect original data models to derivative models that extend them.

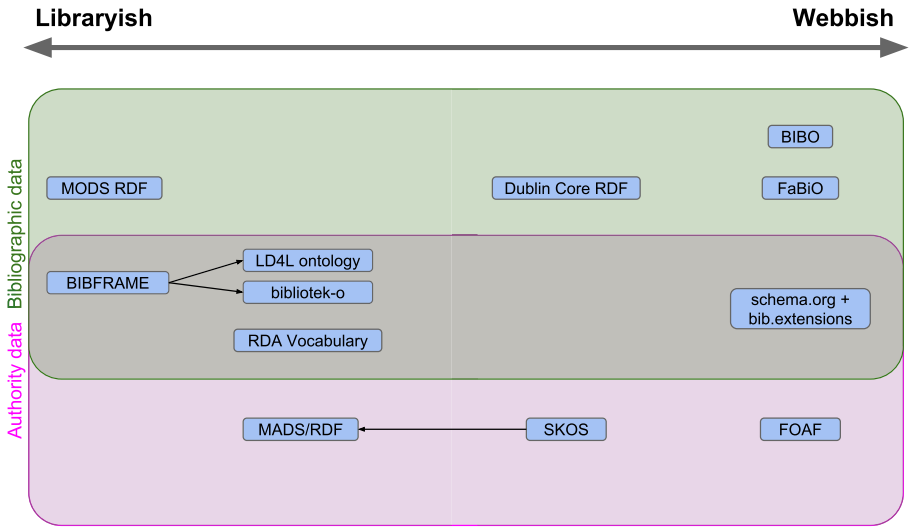


*Figure 2: RDF data models by use case*

FOAF was one of the first Semantic Web ontologies and provides the means to represent people and organizations as real world objects. It has been used in a very large number of Linked Data sets and can be considered extremely *webbish*. SKOS[61] is also a widely used data model that can be used to represent thesauri, classifications and other kinds of controlled vocabularies. However, SKOS represents conceptualizations, not real world objects. MADS/RDF[62] is a data model for authority data that extends SKOS in ways that are important to the library community, which brings it closer to the *libraryish* use case.

Of the data models that are limited to bibliographic data, MODS RDF is best suited for the *libraryish* use case because it covers a lot of the detailed information available in MARC bibliographic records. It is also self-contained, reusing only MADS/RDF for representing authority records. DC is much more oriented towards the *webbish* use case, as it was intended for representing metadata on the web. BIBO and FaBiO are extremely *webbish*, since they build on top of other established RDF data models, including DC, FOAF and SKOS.

61 „SKOS Simple Knowledge Organization System - Home Page," W3C, accessed May 24, 2017, https://www.w3.org/2004/02/skos/.

62 „MADS/RDF Primer," Library of Congress, accessed May 24, 2017, http://www.loc.gov/standards/mads/rdf/.

Among the RDF data models that can represent both bibliographic and authority data, BIBFRAME is closest to the *libraryish* use case, since it is self-contained and able to represent a large amount of detail extracted from MARC records. The LD4L and bibliotek-o ontologies, by aligning or even replacing parts of BIBFRAME with more established RDF data models, bring the data model closer to the *webbish* use case. The RDA vocabularies are also oriented around original representation of bibliographic data as RDF. The Schema.org data model, in contrast, is heavily oriented towards the *webbish* use case, even though it doesn't attempt to build on other established RDF data models.

## 3.4. Choosing a data model

When choosing a data model for bibliographic data among the current options, there are two key questions that affect the decision:

1. Is there a need to represent Works as separate entities, or is it enough to express individual records without grouping them by Work?
2. Is the data going to be maintained natively as RDF (*libraryish* use case), or just shared with the world as Linked Data (*webbish* use case)?

Converting existing data (i.e. MARC) into a modern entity-based model is difficult and may hamper adoption of such data models in practice for real data. All FRBR-based models require extraction of WEMI entities, which requires sophisticated methods and may thus be costly and difficult. For some recent work on FRBR entity extraction, see Pfeifer et al.[63] and Candela et al.[64] as well as a review[65] of "FRBRization" techniques by Decourselle et al. BIBFRAME is somewhat easier because of its more relaxed view about Works, but it still requires remodelling of existing data. Flat, record based data models may thus be more practical at least in the short term.

For maintaining the data natively as RDF, suitable data models include BIBFRAME and the RDA Vocabulary, but it remains largely untested whether they are mature enough for that purpose. The BIBFLOW[66] and LD4P projects are pioneers in this area. For publishing bibliographic data, there are already a lot of data models and reusing one of them would be preferable to adding a new model into the mix.

63 Barbara Pfeifer and Renate Polak-Bennemann, „Zusammenführen was zusammengehört – Intellektuelle und automatische Erfassung von Werken nach RDA," *o-bib. Das offene Bibliotheksjournal* 3.4 (2016): 144-155, http://doi.org/10.5282/o-bib/2016H4S144-155.

64 Gustavo Candela, Pilar Escobar, Rafael Carrasco and Manuel Marco, „Migration of a Library Catalogue into RDA Linked Open Data". *Semantic Web* (2017) (Preprint): 1-11, accessed May 24, 2017, http://www.semantic-web-journal.net/system/files/swj1453.pdf.

65 Joffrey Decourselle, Fabien Duchateau and Nicolas Lumineau, „A Survey of FRBRization Techniques," in *Research and Advanced Technology for Digital Libraries. 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015, Proceedings, ed. Sarantos Kapidakis, Cezary Mazurek and Marcin Werla* (Cham: Springer International Publishing, 2015), 185-196, https://doi.org/10.1007/978-3-319-24592-8.

66 „BIBFLOW," University of California, accessed May 24, 2017, https://bibflow.library.ucdavis.edu/.

 

## 4. How to counter the proliferation

Based on the survey in Section 2 it is clear that libraries which publish Linked Data have all defined their own data models, whether by defining their own schema/ontology or an application profile that reuses other data models such as DC and BIBO. This makes interoperability between bibliographic Linked Data sets difficult. To improve the situation, we suggest some recommendations for future library-oriented Linked Data projects.

### 4.1. Avoiding new data models

If interoperability of bibliographic Linked Data is a goal, then the defining of new data models needs to stop. Libraries should work together to find as much common ground on data models as possible. At the very least, defining application profiles that reuse established data models, as many libraries have done, should be preferred over creating completely new data models from scratch. However, the application profiles that various libraries have already defined are different enough that the interoperability of their respective data sets is currently very limited.

Hillmann, Dunsire and Phipps advocate the creation of highly granular data models based on local requirements, in a bottom-up approach, and then mapping them to shared vocabularies such as DC or the RDA vocabularies.[67] However, this is difficult in practice, for several reasons: First, there is a large difference in the granularity of the various data models that have been used for bibliographic Linked Data. It is not clear how to map a FRBR-based data model to a flat, record based model, or vice versa. Second, the expressivity of RDF Schema, the language most commonly used to define mappings between RDF vocabularies, is limited to simple subclass and subproperty relationships. Third, even when such mappings have been defined, the tools used to harvest and process Linked Data often do not make use of them, since even simple RDFS inference is resource-intensive and error prone.

### 4.2. Improving existing data models

It is clear that the currently available data models do not meet all the needs of libraries as otherwise one of them would already have emerged as the preferred model for bibliographic Linked Data. Libraries planning to develop their Linked Data offerings should get involved with the respective communities to further improve the data models so that they suit their local needs better. It would help if the collaboration around data models was more open, transparent and organized. As an example of this kind of collaboration, libraries in German-speaking countries have formed a working group for coordinating and eventually harmonizing their RDF data models.[68] One way of easing collaboration is to make use of social development platforms such as GitHub for gathering feedback and change suggestions from the larger community.[69] This approach is taken by, among others, the BNE ontology, the RDA Vocabularies, LD4L and Schema.org.

67 Hillmann, Dunsire and Phipps, „Maps and Gaps“.
68 AG KIM Gruppe Titeldaten DINI, „Empfehlungen zur RDF-Repräsentation bibliografischer Daten,“ 2014, accessed May 24, 2017, http://edoc.hu-berlin.de/docviews/abstract.php?id=40697.
69 Hillmann, Dunsire and Phipps, „Maps and Gaps“.

### 4.3. Externally imposed data models

A possible scenario for the future, especially if the library community fails to come up with a common data model, is that a powerful external actor, such as Facebook or one of the major Web search engines, starts harvesting bibliographic data from libraries *en masse*. The harvesting organization would define the exact representation that libraries must use if they don't want to be left out. If the service that is based on this harvested data is attractive enough for the libraries, they would have no choice but to provide their bibliographic data using the externally imposed data model, regardless of how difficult this may be for them and how much data quality will suffer in the conversion. At present this scenario may seem unlikely, but a similar process is already happening for scientific data sets: Google has defined a Schema.org-based model for data sets[70] that publishers of scientific data need to follow if they want their data to be available in future specialized search engines for research data discovery.

## 5. Opening up Finnish bibliographic data

Libraries should have a practical understanding of how they want to produce and use Linked Data for various emerging purposes in the future. They should reshape their data models to enable new partnerships in the production of descriptive metadata. Currently in Finland, new partnerships are being created with museums, archives and the public administration sector alongside the more traditional cooperation with publishers.

At the National Library of Finland we are currently in the process of opening up bibliographic metadata as Linked Open Data, including the national bibliography Fennica, the national discography Viola and the article reference database Arto. After carefully studying different solutions found by other libraries, we chose the Schema.org model for publishing our bibliographic Linked Data. We will represent works extracted from the MARC records as an additional layer, similar to WorldCat Works. A conversion infrastructure based on the reuse of existing conversion tools is currently being constructed.[71] We hope that this decision will make our bibliographic data more interoperable, accessible and understandable, especially for developers and for Web search engines. We also think that for the future of libraries, it is absolutely essential to take every possible effort in order to find a shared vision and a common data model for producing descriptive metadata sustainably and effectively in the era of Linked Data.

## Bibliography


- Gustavo Candela, Pilar Escobar, Rafael Carrasco and Manuel Marco. „Migration of a Library Catalogue into RDA Linked Open Data." *Semantic Web* (2017) (Preprint): 1-11. Accessed 24 May, 2017. http://www.semantic-web-journal.net/system/files/swj1453.pdf.


70 „Science Datasets", Google, accessed May 24, 2017, https://developers.google.com/search/docs/data-types/datasets.

71 „Scripts and configuration for converting MARC bibliographic records into RDF", accessed May 24, 2017, https://github.com/NatLibFi/bib-rdf-pipeline.

 

- Decourselle, Joffrey, Fabien Duchateau and Nicolas Lumineau. „A Survey of FRBRization Techniques." In *Research and Advanced Technology for Digital Libraries. 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015, Proceedings, edited by Sarantos Kapidakis, Cezary Mazurek and Marcin Werla*,185-196. Cham: Springer International Publishing, 2015. https://doi.org/10.1007/978-3-319-24592-8.
- Deliot, Corine. „Publishing the British National Bibliography as Linked Open Data." *Catalogue & Index* 174 (2014): 13-18.
- Harlow, Christina. „Data Munging Tools in Preparation for RDF: Catmandu and LODRefine." *Code4Lib Journal* 30 (2015). Accessed May 24, 2017. http://journal.code4lib.org/articles/11013.
- Hillmann, Diane I., Jon Phipps and Gordon Dunsire. „Maps and gaps: Strategies for Vocabulary Design and Development." Paper presented at the International Conference on Dublin Core and Metadata Applications, Lisbon, Portugal, September 2-6, 2013. Accessed May 24, 2017, http://dcpapers.dublincore.org/pubs/article/view/3673/1896.
- Isaac, Antoine, William Waites, Jeff Young and Marcia Zeng. „Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets." In *W3C Incubator Group Report* (2011). Accessed May 24, 2017, http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/.
- Niininen, Satu, Susanna Nykyri and Osma Suominen. „The Future of Metadata: Open, Linked, and Multilingual – the YSO Case." *Journal of Documentation* 73.3 (2017): 451-465, http://dx.doi.org/10.1108/JD-06-2016-0084.
- Pfeifer, Barbara and Renate Polak-Bennemann. „Zusammenführen was zusammengehört – Intellektuelle und automatische Erfassung von Werken nach RDA." *o-bib. Das offene Bibliotheksjournal* 3, Nr. 4 (2016): 144-155. http://doi.org/10.5282/o-bib/2016H4S144-155.
- Suominen, Osma, Sini Pessala, Jouni Tuominen, Mikko Lappalainen, Susanna Nykyri, Henri Ylikotila, Matias Frosterus and Eero Hyvönen (2014). „Deploying National Ontology Services: From ONKI to Finto." Paper presented at the International Semantic Web Conference (Industry Track), Riva del Garda, Italy, October 19-23, 2014. Accessed May 24, 2017, http://www.ceur-ws.org/Vol-1383/paper6.pdf.
- Yasumatsu, Saho, Akiko Hashizume and Julie Fukuyama. „National Diet Library Dublin Core Metadata Description (DC-NDL): Describing Japanese Metadata and Connecting Pieces of Data." Paper presented at the International Conference on Dublin Core and Metadata Applications, Copenhagen, Denmark, October 13-16, 2016. Accessed May 24, 2017. http://dcevents.dublincore.org/IntConf/dc-2016/paper/view/437/485.